知识管理论坛 ISSN 2095-5472 CN11-6036/C
Knowledge Management Forum

E-mail: kmf@mail.las.ac.cn
http://www.kmf.ac.cn

【学术探索】

国际数据管护的科学知识图谱研究

◎ 虞晨琳 [1,2]
[1] 中国科学院文献情报中心 北京 100190 [2] 中国科学院大学 北京 100049

摘要：[**目的 / 意义**] 数据管护是信息化科研环境下研究数据管理的重要部分，梳理国际已有相关研究成果，以期全面认识数据管护，为国内数据管理研究提供参考。[**方法 / 过程**] 以 Web of Science 为数据源，检索时间截至 2016 年 10 月，检索词为数据管护，将检索到的文献作为研究对象，基于文献共现和共被引分析方法，利用 CiteSpace Ⅲ软件工具，绘制国际数据管护的知识图谱，采用内容分析法，基于研究的学科分布、研究机构、研究人员以及知识基础的 4 个维度，对国际数据管护研究进行解读、分析与总结。[**结果 / 结论**] 国际数据管护研究始于 2000 年，已经步入成熟期，并形成特定的研究学科、机构和群体，研究的知识基础主要为数据描述、集成与关联、科研过程的数据维护和增值活动、数据管护利益相关者以及图书馆服务新模式。

关键词：数据管护　数据管理　研究数据　知识图谱
分类号：G250

引用格式：虞晨琳 . 国际数据管护的科学知识图谱研究 [J/OL]. 知识管理论坛 , 2017, 2(3): 201-213[引用日期]. http://www.kmf.ac.cn/p/1/137/.

1 引言

随着 E-Science 的发展，科研行为的主要特征是基于数据的科学探索，研究数据是科研活动的驱动力，科学研究已步入以数据密集型为特征的大数据科研范式 [1]。大数据时代，研究数据的内涵与特点发生改变，其来源范围广、类型多样、数据体量巨大以及数据流实时变化，被称之为科学大数据 [2]。因此，以往的数据管理模式因不能适应研究数据的管理，而使得研究数据易遭到损坏与污染，数据不能得到有效利用和长久保存，影响现阶段的科学研究行为的进行。各领域学者基于自身学术背景对研究数据管护 (data curtain, DC）进行了理论研究与实践探索。笔者将对国际学术界的数据管护研究进行梳理，以期整体、全面地认识与把握数据管护研究的整体面貌。

2 数据管护定义

英国数据管护中心 (Digital Curation Centre, DCC) 对数据管护进行明确定义：数据管护是指贯穿数字化研究数据整个生命周期的维护、保存和增值的动态主动的管理活动；对研究数据进行主动的管理，其目的是为了确保数据在未来研究价值的威胁、降低数字老化的风险；置于可信的数字化存储库中的管护数据，可促进英国研究领域的数据共享；数据管护可减少数

作者简介：虞晨琳 (ORCID：0000-0002-7925-6548)，硕士研究生，E-mail: yuchenlin@mail.las.ac.cn
收稿日期：2017-02-20　发表日期：2017-06-16　本文责任编辑：王传清

据创建的重复工作，并通过增强高质量研究的可用性来提高数据的长期价值 [3]。联合信息系统委员会 (Joint Information Systems Committee, JISC) 指出 , 数据管护是在数字数据和研究成果的整个生命周期内 , 维护和利用它们以服务当前和未来的用户的一系列活动 [4]。

从档案视角解读，认为数据管护是将数字保存、数字图书馆管理、数字归档和数据管理阶段性介入活动进行融合成一个整体；数据管护实质是贯穿整个数据生命周期的管护活动，数据管护术语的产生，由于数字归档的含义在信息资源保存领域的滥用，使得数字归档的含义遭到曲解，使得数字资源的长期、全过程管理的研究需要创建新的术语来准确描述数字资源的生命周期管理的研究 [5]。

美国伊利诺伊大学图书馆与信息科学学院提出数据管护是在学术研究、科学和教育活动中主动、持续地贯穿数据生命周期的数据管理活动，通过数据认证、归档、管理、保存和描述来促进数据的检索发现、长期保存和增值重用 [6]。

综上所述，数据管护具有以下特点：①数据管护是一种主动、持续和不间断的数据管理，贯穿整个研究数据的生命周期，确保研究数据管理过程是一条可追溯的连续链条；②数据管护目的是维护和增值研究数据的价值，确保数据的真实可靠和长期可用，满足现在和未来的使用需求；③数据管护促进研究数据资源的检索与发现、共享与利用、减少科研资源的重复建设。

3 研究结果分析

3.1 数据与方法

为全面把握国际数据管护研究情况，避免遗漏重要文献，本文所选取的统计数据来源于 Web of Science （WOS）核心合集数据库，以 “digital curation” “data curation” 为主题或标题进行检索，时间跨度：1900-2016 年，文献类型： 包 括“article，editorial，letter，proceeding paper，review”5 类，检索时间为 2016 年 10 月 31 日，并对检索结果进行去重、清洗，最终得到 319 条文献记录。

国外数据管护研究的文献增长趋势符合普赖斯提出的科学文献指数增长的普遍规律，拟合优度 R^2 为 0.974（见图 1）。国外数据管护研究始于 2000 年，2000-2005 年间的发文量少，发展极为缓慢，研究处于起步阶段；2006-2013 年间的年发文量呈现增长态势，实际发文量都超过理论值，研究处于快速增长期; 2013 年之后，实际发文量小于理论值，且两者之间的差距逐年拉大，研究步入成熟期。数据管护的年发文量呈绝对值持续增长趋势，自 2013 年起，每年发文量均在 40 篇以上，2015 年达到 62 篇。

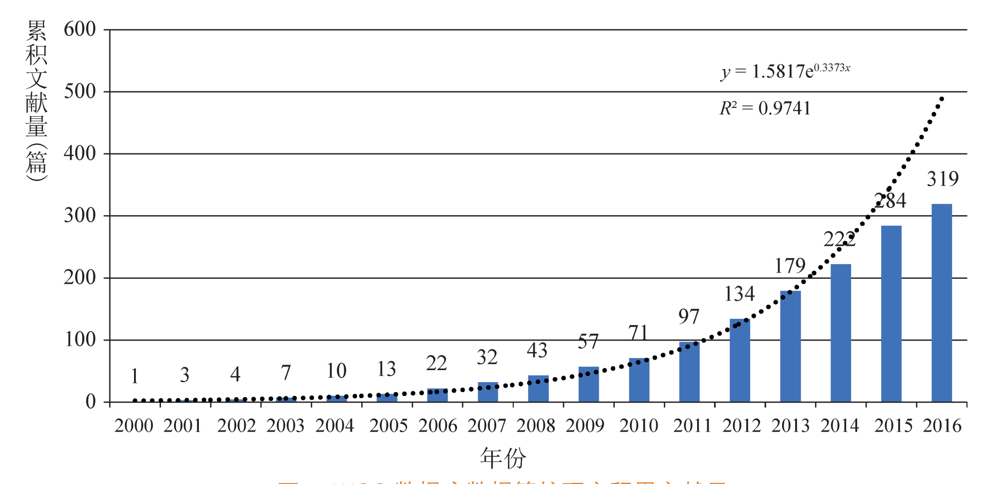


图 1 WOS 数据库数据管护研究积累文献量

知识管理论坛，2017(3):201–213
DOI：10.13266/j.issn.2095–5472.2017.022

本文所选取的研究方法是科学知识图谱，科学知识图谱是将信息可视化技术、应用数学、图形学、计算机科学等与科学计量学结合起来的交叉科学研究方法，可将科学前沿领域的海量文献数据信息转换为可视化图像，展示单凭个人经验难以直观获得的学科前沿领域的总体图景、发展态势与结构特征。具体分析方法是基于共现分析法来明确国外数据管护的研究主体；利用共被引分析展现国外数据管护的知识基础。

3.2 数据管护的研究主体

利用 CiteSpace 软件共现图谱分析法，从学科分布、研究机构、作者分析 3 个维度对施引文献进行分析，以探求数据管护的研究主体。

3.2.1 学科分布分析科学知识图谱

如图 2 所示，计算机科学与图书情报学的节点年轮较大，表明学科的发文数量多；节点年轮颜色由蓝、绿、黄组成，暗示研究跨 3 个时间段，长期时间关注且持续性研究。生物化学研究方法、天文与天体物理、计算机科学、成像科学与照相技术、统计与概率、地理学、生物化学与分子生物、遥感、基因与遗传学等学科的节点被紫圈标注出来，代表节点具有较大的中心度（不小于 0.1），处于在网络结构中重要的中心位置，在研究中具有重要影响力。

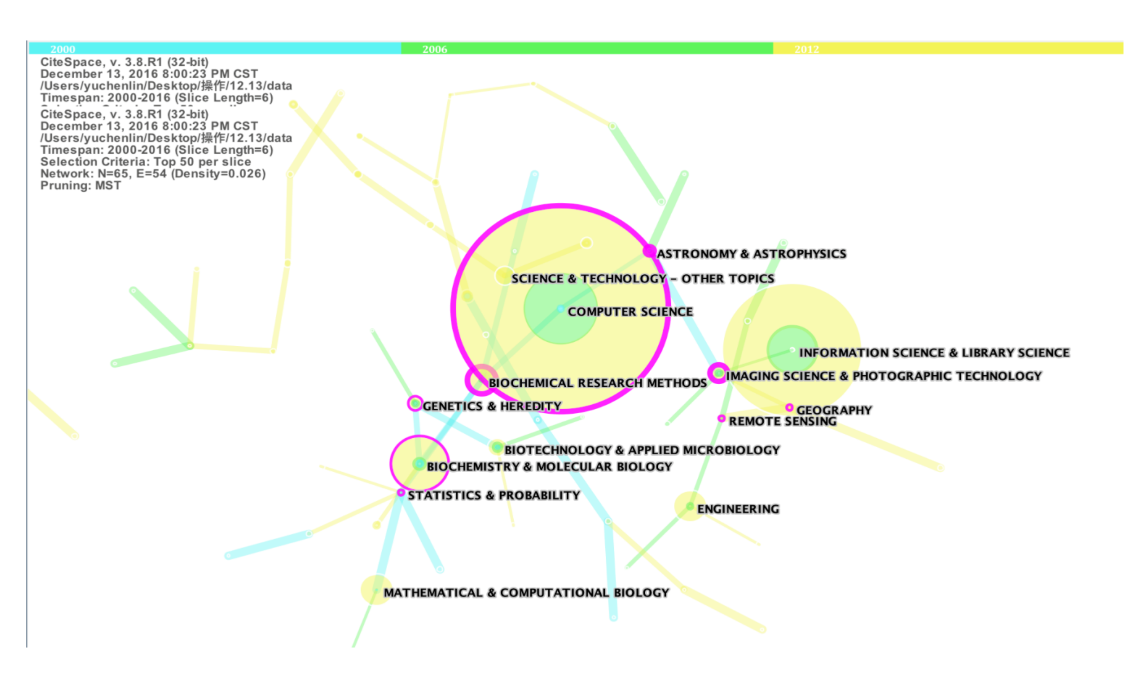


图 2 WOS 数据库数据管护的学科分布

从学科分布来看，数据管护研究具有多学科性，应用学科和基础学科均关注数据管护方面问题，积极开展相应的研究工作，产生这种现象的原因主要为：①研究数据主要由具体的基础学科产生。研究数据来源于科学研究的观测、探测、调查和综合分析所获得的数值型的事实记录，随着 21 世纪的信息技术革命，新一代科学研究的手段与方式的应用，促使研究数据的生产方式步入自动式化感知式系统阶段。研究数据具有学科背景属性，基础学科多围绕学科的特定项目开展数据管护研究，以满足自身学科知识体系对研究数据的管护的特定需要。②不同学科的研究数据在管理与服务具有共同属性。应用学科夯实了数字化科研的基础以及统一了研究数据的技术标准，这些称为了数据管护中的网络基础设施的依托、信息技术的支撑、政策指导与管护理论的提供了强有力的支持。



知识管理论坛
2017 年 第 3 期(总第 9 期)

计算机科学在数据管护的研究方向主要是人工智能、信息系统、跨学科应用、软件工程与理论方法，从全方面对数据管护研究进行技术支持，其研究始于 2001 年。生命科学与生物医学对数据管护研究力度与重视程度不亚于计算机科学，随着新一代测序工具与技术出现，基因研究产生海量的基因数据，因此，生命科学与生物医学对于基因数据管理需求增大，需要确保基因数据的及时更新、实时维护、关联和集成资源、长期保存与有效获取等，驱动科学研究的新发现。图书情报学的发文数高达 84 篇，科学体量较大，学术影响力较强，是推动数据管护研究进展的主力军之一。

3.2.2 研究机构分析

由图 3 可见，北卡罗来纳大学教堂山分校、爱丁堡大学、普渡大学、格拉斯哥大学、约翰·霍普金斯大学、南佛罗里达大学以及圣迭戈加利福尼亚大学在数据管护研究上比较活跃。

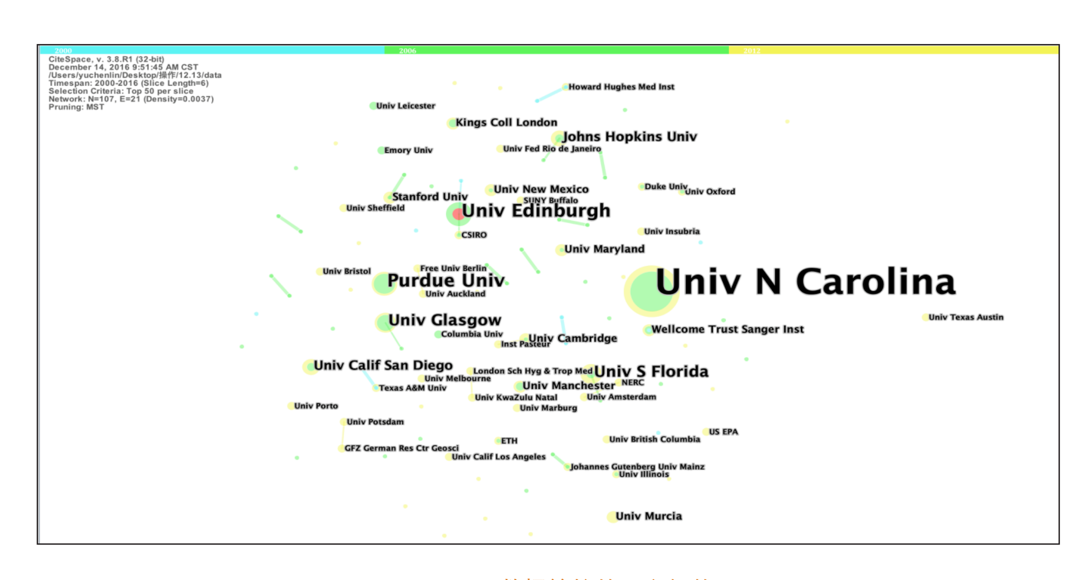


图 3 WOS 数据管护的研究机构

突现是指变量值在短时间内发生很大变化，突现信息是一种可用来度量更深层变化的手段，对机构突现的研究，能够把握机构在数据管护研究上的关键转变节点。北卡罗来纳大学教堂山分校 2007 年共有 4 篇关于数据管护的文献，主要为数据管护的人才培养和软件工具研发的研究。其图书馆与信息科学学院承担的数据管护课程 (Digital Curation Curriculum，DigCCurr) 项目，包括培育数据管护的研究生层次专业人才，探索数据管护课程设置 [7]；界定数据管护人才以及数据管护应具备技能与知识 [8]。The Vidarch Project1 项目捕获数据资源的相关信息，基于数据资源的元数据和上下文本信息关系，实现数据资源的全面注释 [9]；研发 ContextMiner 2 工具，帮助数据管护人在数据库中进行数据查询、编译及存储 [10]。爱丁堡大学 2004-2007 年共有 4 篇关于数据管护的文献。面对生物数据爆发式增长，P. Buneman 倡议对数据库进行管护，确保数据的安全可靠 [11]；P. Buneman 同时阐释数据管护的两种不同的文化，档案专家、管护者侧重对数据资源的长期保存与可靠访问，研究者侧重数据资源的可视化、注释与关联 [12]；C. Rusbridge 等认为 DCC 成立将更好地指导数据管护活动的开展 [13]；M. McGinley 呼吁将数据管护纳入法律层面，以此将有效地指导研究数据的开放或保密 [14]。普渡大学在 2008 年发表 2 篇关于数据管护文献。普渡大学图书馆在图书馆学和档案学原理的指导下，利用分布式机构知识库设施基础，开展具体学科的研究数据管理的探索，为数据管护研

知识管理论坛，2017(3):201–213
DOI: 10.13266/j.issn.2095-5472.2017.022

究提供实践案例[15]；M. Y. Eltabakh 研发生物数据库的可扩展数据库引擎，支持研究者对生物数据库系统进行统一的数据管理，如数据及派生信息的注释、存储、数据查询和跟踪等，促进普渡大学的研究数据管理[16]。

3.2.3 作者分析

如图 4 所示，节点年轮的颜色变化反映了研究者的活跃时段，笔者依据图谱的时间分区的颜色变化，将数据管护研究领域的主要研究者分为三代研究者，以 2006 年和 2012 年作为时间分区的分界点。

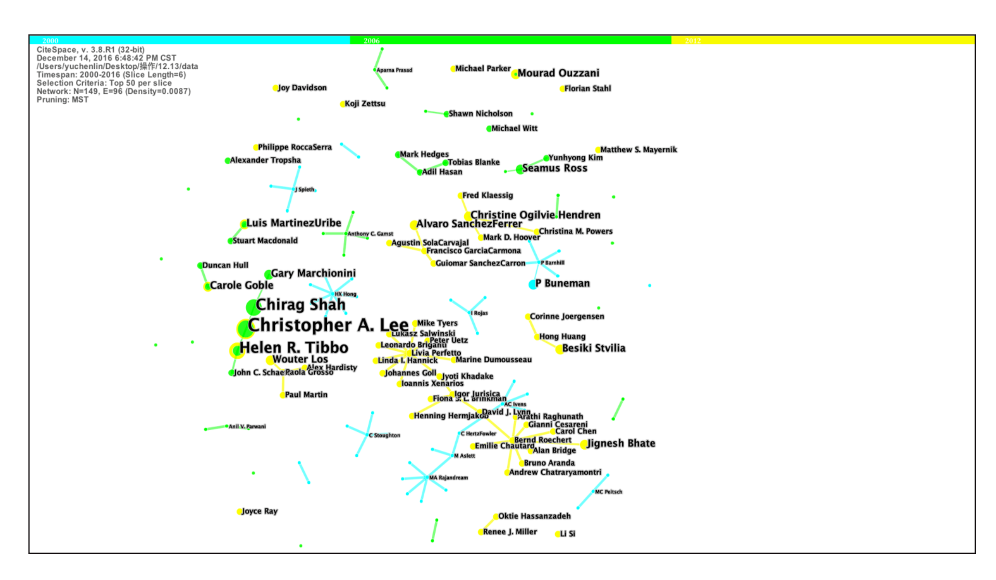


图 4 WOS 数据管护的作者分析

第一代研究者的节点以蓝色为主，随着科研信息化的展开，研究者对研究数据管护的需求不断增加。P. Buneman 团队倡议及阐述数据管护以及数据管护中心成立的意义；P. Martin 团队研发基因数据库的集成分析工具，支持数据集成化研究。第二代研究者的节点以绿色为主，主要是图情及计算机领域围绕研究数据管护展开的研究活动。C. Prom 团队从数据管护教育角度，主持开展数据管护课程（DigCCurr）和数据管护差距弥补课程（Closingthe Digital Curation Gap）以储备数据管护的专业人才；L. Martinez-Uribe 团队研究图书馆在数据管护的角色定位、服务创新；S. Ross 团队研发文本流派分类方法自动获取元数据。第三代研究者的节点以黄色为主，研究主要是针对特定学科开展的细粒度的数据管护活动，Á. Sánchez-Ferrer 团队基于生物基因需求，提出数据管护的具体要求；W. Los 团队建立数据管护以此来推进数据资源共享开放；C. Jandrasits 团队从纳米领域提出数据管护的重要性；B. Stvilia 团队从基因领域出发，研究数据管护以及数据质量要求；J. Bhate 团队介绍国际分子交换联盟中心（IMEx Central）实施交互质量控制、交叉管护等数据管护措施。

3.3 数据管护研究的知识基础

由图 5 可知，文献共被引网络主要为 8 个聚类。基于被引文献和施引文献、聚类标签对各类的研究内容和核心观点进行解读，发现研究内容大致可分为数据管护对科研活动的新价值、数据管护的软硬件设施的建设、数据管护在具体学科的应用、数据管护的利益相关者以及图书馆的服务模式几方面。

3.3.1 数据管护对科研活动的新价值

表 1 列出聚类 3#scientific data 的被引文献和施引文献，阐释科学数据对科研活动的新价值，这些文献主要研究了如何使用数据管护实现对数据的维护和增值，涉及到科研工作流程、数据共享及出版的管理。科学研究具有数

据驱动性和开放协作性，数据共享可以支持科学研究的再现或验证，确保研究结果为公众所用，方便其他人利用现有数据开展新研究，提升研究创新水平 [17]。

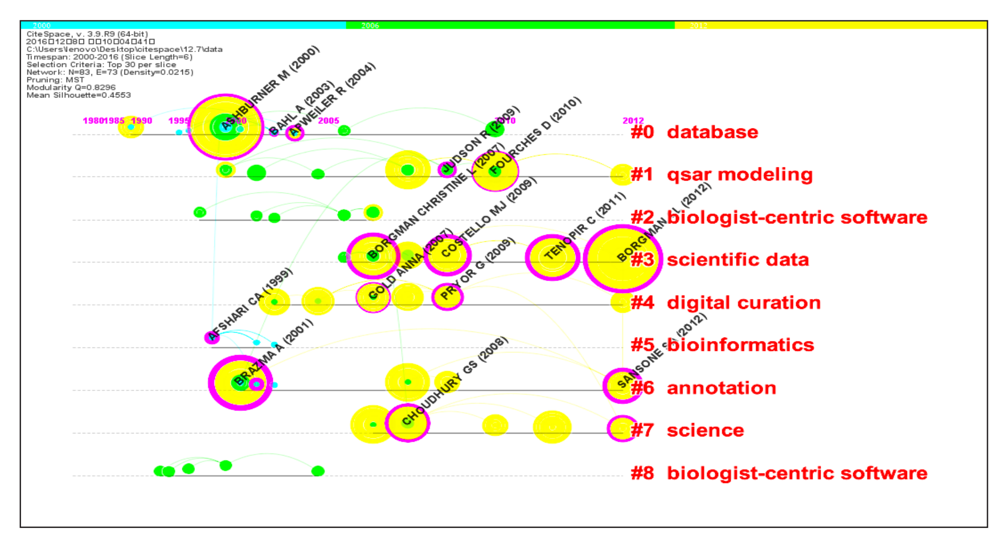


图 5 数据管护研究文献共被引时间线程图

表 1 聚类 3#scientific data 核心被引文献与施引文献

被引文献					施引文献	
被引	中心度	作者	年份	文献标题	引用	文献标题
11	0.48	C. L. Borgman	2012	*The conundrum of sharing research data*	[2]	*Data sharing, small science and institutional repositories*
7	0.45	C. Tenopir	2011	*Data Sharing by Scientists: Practices and Perceptions*	[4]	*Designing submission and workflow services for preserving interdisciplinary scientific data*
6	0.45	M. J. Costello	2009	*Motivating Online Publication of Data*		
7	0.44	C. L. Borgman	2007	*Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries*		

科学界对小研究数据潜在价值的认识加深 [18]，P. Borgman 以栖息地生态学为例，介绍了数字图书馆利用嵌入式网络感知中心，来支持“小科学”学科的数据管理，以便解决小研究数据趋向于异质、个人管理的状态或是未被保存、未被管理的状态 [47]。尽管海量研究数据产生，使得数据洪流现象出现，但只有少数领域出现数据共享，C. Tenopir 等 2011 年对 1 329 名科学家进行数据共享实践与理论调研，发现阻碍科学家进行数据共享首要原因是时间不足和资金缺乏，其次是开放平台、标准规范、政策制定等 [19]。M. H. Cragin 等承担的 Data Curation Profiles 项目是基于研究者角度对数据共享问题进行研究，从分享什么数据、何时和与谁分享的 3 个维度分析研究者数据共享行为 [20]；P. Borgman 分析什么数据应该被共享、被谁共享、在什么条件下共享、为什么共享以及要做什么努力等方面，能帮助认识数据共享；以上研究为数据政策制定和数据实践开展提供了指导 [17]。

知识管理论坛，2017(3):201-213
DOI: 10.13266/j.issn.2095-5472.2017.022

M. J. Costello 提出以数据出版代替数据共享，构建数据的引用与访问系统，激励环境、生物学科学家发布研究数据，解决数据可用性问题 [21]。R. R. Downs 和 R. S. Chen. 设计跨学科数据提交的工作流，便于满足跨领域研究的科研人员提交数据的需求 [22]。

3.3.2 数据管护的软硬件设施建设

数据管护的软硬件设施建设包括支撑数据管护的平台的基础设施，支持数据集成和关联的软件技术。表 2 列出聚类 2#biologist-centricsoftware 的被引文献和施引文献是面向数据管护的基础设施的建设研究，这些文献主要是探讨支撑管护软件研发和平台构建、服务体系建设以及最佳实践探索。

表 2 聚类 2#biologist-centricsoftware 核心被引文献与施引文献

被引文献					施引文献	
被引	中心度	作者	年份	文献标题	引用	文献标题
2	0.02	C. Lagoze	2006	*Fedora: an architecture for complex objects and their relationships*	[2]	*Receptivity to Library Involvement in Scientific Data Curation: A Case Study at the University of Colorado Boulder*
3	0	D. S. Brandt	2007	*Librarians as partners in e-research: Purdue University Libraries promote collaboration*	[6]	*Curation and preservation of research data in an iRODS data grid*
3	0	L. Lyon	2007	*Dealing with Data: Roles, Rights, Responsibilities and Relationships*		

开源数字仓储软件 (Fedora) 描述数字对象及之间的复杂关系，为组织机构在管理及保存数字资源方面提供基础 [23]。iRODS（integrated Rule-Oriented Data System）的数据网格帮助用户高效、简易管理各类数据资源 [24]。英国图书馆与信息网络办公室总结数据管护的服务框架，鉴定关键利益主体，分析其责任、权利与协作方式，确定数据管理的目标（数据的保存、访问和重用），确定实现目标的机制、流程和实践 [25]。普渡大学图书馆在 e-Science 环境下，构建面向科研的嵌入式服务的协同结构，开展研究数据管理服务，包括数据描述、类型和格式的标准、收集、组织、归档与保存 [26]；科罗拉多大学博尔德分校图书馆参与领域科学的数据管护的过程，表明图书馆在专业人才、基础设施与信息服务的优势将有助于开展数据管护活动 [27]。以上图书馆的探索成为数据管护的最佳实践。

表 3 列出聚类 6#annotation 的被引文献和施引文献是基于数据集成和关联的数据管护，通过构建大规模知识化的科学数据网络，便于研究者深入挖掘和有效解释科研数据中各类资源对象的内涵和关系。

基因芯片数据协会组织开发了微阵列数据标准，规范了微阵列实验解释的最小信息描述 [28]，促进国际上基因组学的实验室及公共数据库的数据交流。C. A. Ball 评述微阵列数据标准，规范了微阵列实验数据的注释描述和交换标准，辅助微阵列数据库的建设和数据分析工具的开发，促使高质量的基因表达数据的共享，为基因研究的标准化铺平道路 [29]。S. A. Sansone 提出以技术手段和奖励机制促进生物数据的互操作性，以提高科学社群对研究数据的充分利用和开放共享 [30]。D. Howe 认为生物研究数据管理和生物学数据管理的出现，解决不断增长的高质量数据需求与有限、落后的数据管理之间的矛盾 [31]。B. M. Good 等通过语义维基构建生物医学的语义网链接，直接嵌入维基百科编辑器来计算文章上下文的语义关系，增

强维基百科文章的语义呈现，便于用户查询与发现[32]。

表 3 聚类 6#annotation 的被引文献和施引文献

被引文献					施引文献	
被引	中心度	作者	年份	文献标题	引用	文献标题
8	0.61	A. Brazma	2001	*Minimum information about a microarray experiment (MIAME)—toward standards for microarray data*	[7]	*ArrayTrack - Supporting toxicogenomic research at the US Food and Drug Administration national Center for Toxicological Research*
5	0.51	S. A. Sansone	2012	*Toward interoperable bioscience data*	[2]	*Building a biomedical semantic network in Wikipedia with Semantic Wiki Links*
1	0.47	C. A. Ball	2002	*Standards for Microarray Data*		
7	0.03	D. Howe	2008	*Big data: The future of biocuration*		

3.3.3 数据管护在具体学科的应用

数据管护在生物学科、化学信息学与生物信息学方面得到充分运用。表 4 列出聚类 0#database 的被引文献和施引文献是数据管护在生物学科的具体应用，这些文献主要是基于领域本体与元数据的数据描述的管护活动，为生物数据的描述和分类实现格式化，为计算机处理创造可能。

表 4 聚类 0#database 核心被引文献与施引文献

被引文献					施引文献	
被引	中心度	作者	年份	文献标题	引用	文献标题
11	0.29	M. Ashburner	2000	*Gene Ontology: tool for the unification of biology*	[6]	*Integration of tools and resources for display and analysis of genomic data for protozoan parasites*
2	0.21	R. Apweiler	2004	*UniProt: the Universal Protein knowledgebas*	[2]	*Long-term genome-wide blood RNA expression profiles yield novel molecular response candidates for IFN-beta-1b treatment in relapsing remitting*
1	0.19	A. Bahl	2003	*PlasmoDB: the Plasmodium genome resource. A database integrating experimental and computational data*	[1]	*Integration of data in biosciences*

随着新一代基因测序技术的快速发展，使得基因组和转录组开始进入高通量测序，实验室和基因数据库得到海量核序列数，但是对核序列数的描述和保存格式不统一，严重阻碍了学术交流与资源共享。基因本体的出现，统一了规范基因功能注释和描述[33]；生命研究数据库采用基因本体来对研究数据进行标注，通用蛋白质资源数据库（UniProt）为科学社群提供集成、高质量、可获取的蛋白质资源数据[34]，PlasmoDB 数据库通过疟原

ChinaXiv合作期刊
知识管理论坛，2017(3):201-213
DOI：10.13266/j.issn.2095-5472.2017.022

虫基因注释标准化，关联基因组定位、转录本信息等各种信息，方便疟疾研究者查询[35]。数据的描述、注释以及保存格式的规范，有助于研究的新发现，通过统一基因本体术语，便于集成高质量的数据资源，便于发现基因之间的相互作用的证据[36]。

表 5 列出聚类 1#QSARmodeling 的被引文献和施引文献是数据管护在化学信息学的具体应用，这些文献主要是围绕研究数据建模过程的管护活动，依据数学原理，探索数据之间的关系，提取信息及发现知识等。定量构效关系 (quantitative structure activity relationship，QSAR) 作为化学信息学的主要研究方法，是对化合物结构与其活性之间关系的定量描述研究[37]。

表 5 聚类 1#QSAR modeling 核心被引文献与施引文献

被引文献					施引文献	
被引	中心度	作者	年份	文献标题	引用	文献标题
2	0.27	R. Judson	2009	*The toxicity data landscape for environmental chemicals*	[2]	*Genotype-phenotype databases: challenges and solutions for the post-genomic era*
7	0.13	D. Fourches	2010	*Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research*	[2]	*A quality alert and call for improved curation of public chemistry databases*
3	0.1	M. Kanehisa	2000	*KEGG: Kyoto Encyclopedia of Genes and Genomes*	[2]	*Data governance in predictive toxicology: a review*
					[2]	*Best practices for QSAR model development, validation, and exploitation*

建立研究数据的汇聚机制与模型，如集成计算毒理学资源 (Aggregated Computational Toxicology Resource, ACToR)、京都基因和基因组学百科全书 (Kyoto Encyclopedia of Genesand Genomes, KEGG) 和基因型—表现型数据库（Genotype-phenotype databases），以解决数据的多源、异构带来的数据使用效率低的难题。科研信息化的推进，数据驱动科学研究的发展，数据质量直接决定研究的成败。化学数据建模分析过程采用标准规范[38]，划定分析阶段，来确保 QSAR 模型分析结果的有效性[39]。面对预测毒理学的数据的来源涉及学科广、数据的表示灵活多样，F. Xin 认为数据管护能确保预测毒理学的计算基础的数据高质量，推进学科发展[40]。A. J. Williams 和 S.EKINS 倡议化学数据库采用数据管护，来保障数据质量，推动科研进展[41]。

表 6 列出聚类 5#bioinformatics 的被引文献和施引文献是数据管护在生物信息学的具体应用，这些文献论证了数据管护是如何支持生物信息学的研究新模式。J. Bellenson 指出，微阵列芯片技术在鉴定致癌物质与环境危害的应用，促使毒理学研究的范式由假设驱动的研究转向数据驱动的实验[42]，数据对科研的重要性日益显著。W. Tong 等指出 arraytrack 具有集合毒理学的数据存储、分析和可视化的功能，支持毒物学研究的进展与新发现[43]。

3.3.4 数据管护的利益相关者以及图书馆的服务模式

表 7 列出聚类 4#digitalcuration 的被引文献和施引文献确定了数据管护的利益相关者，这些文献主要是围绕数据管护利益相关者展开的角色定位、职责划定和相互协作研究。

表 6 聚类 5#bioinformatics 核心被引文献与施引文献

被引文献					施引文献	
被引	中心度	作者	年份	文献标题	引用	文献标题
1	0.46	C.A. Afshari	1999	*Application of complementary DNA microarray technology to carcinogen identification, toxicology, and drug safety evaluation*	[7]	*ArrayTrack - Supporting toxicogenomic research at the US Food and Drug Administration national Center for Toxicological Research*
1	0.03	J. L. Bellenson	1999	*Expression data and the bioinformatics challenges*		

表 7 聚类 4#digital curation 核心被引文献与施引文献

被引文献					施引文献	
被引	中心度	作者	年份	文献标题	引用	文献标题
4	0.22	G. Pryor	2009	*Skilling up to do data: whose role, whose responsibility, whose career?*	[2]	*Placing the horse before the cart: conceptual and technical dimensions of digital curation*
5	0.13	A. Gold	2007	*Cyberinfrastructure, data, and libraries Libraries and the data challenge: roles and actions for librar*		
5	0.10	National Science Board	2005	*long-lived digital data collections:enabling research and education in the 21st century*		

美国国家科学委员会（National Science Board，NSB）发布《21 世界长期数字数据集合研究与教育》，明确了管理层面对长期数字数据集合管理的重视，开展数据管理研究以及教育培训，以支撑 2000 年以后的科学研究。基于数据在不同阶段的管理要求，提出不同机构、部门的数据服务角色定位，以实现数据管理服务角色的协作，实现数据管理服务的目标 [44]。图书馆作为信息资源管理的参与者，拓展和延伸数据服务，定位管理角色与职责，研究技术标准和数据生命周期理论等，以期在研究数据管理乃至科学研究中发挥重要作用。H. R. Tibbo 从社会科学角度度审视数据管护，尽管数据管护的发展离不开计算机技术的支撑，但社会科学对数据资产的长期管护更具有指导 [45]。

表 8 列出聚类 7#science 的被引文献和施引文献描述了科研新模式下图书馆的探索，这些文献主要是描述了图书馆的数据管护服务模式。L.Lyon 指出，随着“信息转变”，图书馆需要审视在数据驱动科研环境下的机构目标和服务范围 [46]。P. Hswe 和 P Hswe 从学术图书馆在人员配置、基础设施及服务定位角度，论述图书馆参与数据管理的必要性和参与模式，指出图书馆将出现新的职业角色来满足数据管理的需要 [47]。G. S. Choudhury 针对约翰霍普金斯大学已有的机构库等基础设施开展数据管护服务，强调数据科学家和数据人文专家等新角色在数据管护中发挥的作用，能全面支持高校研究数据管理 [48]。L. M.Delserone 论述了明尼苏达大学图书馆与机构库、信息部门等协同合作，共同规划建设学校的数据管护的基础设施；同时图书馆配置专业人才队伍，满足图书馆开展数据管理与服务的要求，建设“科学馆员队伍”[49]。L. Lyon 基于 Research360 的机构研究生命周期模型，总结图书馆开展数据管护服务的 10 个阶段，包括数据管理要求、计划、信息学基础、引用、培训、许可、鉴定、存储、获取、影响 [46]。

知识管理论坛，2017(3):201–213
DOI：10.13266/j.issn.2095–5472.2017.022

表 8 聚类 7#science 的被引文献和施引文献

被引文献					施引文献	
被引	中心度	作者	年份	文献标题	引用	文献标题
6	0.35	G. S. Choudhury	2008	*Case study in data curation at Johns Hopkins University*	[3]	*Data management services in libraries*
6	0.32	L. M. Delserone	2008	*At the watershed: preparing for research data management and stewardship at the University of Minnesota Libraries*		
4	0.24	L. Lyon	2012	*The informatics transform: re-engineering libraries for the data decade*		

4 结语

随着 21 世纪的信息技术革命，科学研究范式向数据密集型转变，共同推动数据管护研究的兴起。对国际的数据管护研究的分析和解读表明，研究主体具有多学科性，其中，生命科学与生物医学基于自身学科知识体系，围绕特定项目进行数据管护的研究；计算机与图情等应用学科则基于研究数据的通性，研究通用的研究数据的基础设施与技术标准规范。研究主体的机构主要集中在欧美，其中北卡罗来纳大学教堂山分校、爱丁堡大学和普渡大学在数据管护领域比较活跃，具有很大影响力。相较国外，中国对数据管护的研究相对薄弱，武汉大学信息管理学院在国际数据管护的专业人才培养上开展深入调研与分析，具有较强的影响力。研究主体的学者合作不够紧密，缺少稳定的、高质量的研究团队。数据管护的知识基础集中于数据管护对科研活动的新价值、数据管护的软硬件设施的建设、数据管护在具体学科的应用、数据管护的利益相关者以及图书馆的服务模式。基于上述对国际数据管护研究的英文文献的梳理，望能为国内开展数据管护研究带来启示与借鉴。

参考文献：

[1] 吴金红 , 陈勇跃 , 胡慕海 . e-Science 环境下科学数据监管中的质量控制模型研究 [J]. 情报学报 , 2016, 35(3): 237-45.
[2] 郭华东 , 王力哲 , 陈方 , 等 . 科学大数据与数字地球 [J]. 科学通报 , 2014 (12): 1047-1054.
[3] What is digital curation [EB/OL]. [2017-04-10]. http://www.dcc.ac.uk/digital-curation/what-digital-curation.
[4] BEAGRIE N, POTHEN P. Digital curation: digital archives, libraries and e-Science seminar [EB/OL]. [2017-04-10].http://www.ariadne.ac.uk/issue30/digital-curation/.
[5] CUNNINGHAM A. Digital curation/digital archiving: a view from the National Archives of Australia [J]. The American archivist, 2008, 71(2): 530-573.
[6] MURAKAMI Y. Metal fatigue: effects of small defects and nonmetallic inclusions [M]. Amsterdam: Elsevier, 2002.
[7] LEE C A, TIBBO H R, SCHAEFER J C. DigCCurr: Building an International Digital Curation Curriculum & the Carolina Digital Curation Fellowship Program[EB/OL]. [2017-04-10]. http://chinesesites.library.ingentaconnect.com/content/ist/ac/2007/00002007/00000001/art00025.
[8] LEE C A, TIBBO H R, SCHAEFER J C. Defining what digital curators do and what they need to know: the DigCCurr project[EB/OL]. [2017-04-10]. http://dl.acm.org/citation.cfm?id=1255183.
[9] SHAH C, MARCHIONINI G. Capturing relevant information for digital curation[EB/OL]. [2017-04-10]. https://ils.unc.edu/vidarch/Shah-JCDL2007poster.pdf.
[10] SHAH C, MARCHIONINI G. ContextMiner: A tool for digital library curators[EB/OL]. [2017-04-10]. https://ils.unc.edu/vidarch/Shah-JCDL2007demo.pdf.
[11] BUNEMAN P, CHENEY J, TAN W C, et al. Curated databases[EB/OL]. [2017-04-10]. http://dl.acm.org/citation.cfm?id=1376918.
[12] BUNEMAN P. The Two Cultures of Digital Curation[EB/OL]. [2017-04-10]. http://www.inf.ed.ac.uk/teaching/courses/ad/lectures04/buneman.pdf.
[13] RUSBRIDGE C, BURNHILL P, ROSS S, et al. The digital curation centre: a vision for digital curation[EB/OL]. [2017-04-10]. http://ieeexplore.ieee.org/abstract/

document/1612461/.
[14] MCGINLEY M. The legal environment of digital curation–a question of balance for the digital librarian[EB/OL]. [2017-04-10]. https://link.springer.com/chapter/10.1007%2F978-3-540-74851-9_62?LI=true.
[15] WITT M. Institutional repositories and research data curation in a distributed environment [J]. Library trends, 2008, 57(2): 191-201.
[16] ELTABAKH M Y, OUZZANI M, AREF W G, et al. Managing biological data using bdbms[EB/OL]. [2017-04-10]. http://ieeexplore.ieee.org/abstract/document/4497631/.
[17] BORGMAN C L. The conundrum of sharing research data[J]. Journal of the American Society for Information Science and Technology, 2012, 63(6): 1059-1078.
[18] BORGMAN C L, WALLIS J C, ENYEDY N. Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries [J]. International journal on digital dibraries, 2007, 7(1/2): 17-30.
[19] TENOPIR C, ALLARD S, DOUGLASS K, et al. Data sharing by scientists: practices and perceptions [J]. PloS one, 2011, 6(6): e21101.
[20] CRAGIN M H, PALMER C L, CARLSON J R, et al. Data sharing, small science and institutional repositories[J]. Philosophical transactions of the Royal Society of London A: mathematical, physical and engineering sciences, 2010, 368(1926): 4023-4038.
[21] COSTELLO M J. Motivating online publication of data [J]. BioScience, 2009, 59(5): 418-427.
[22] DOWNS R R, CHEN R S. Designing submission and workflow services for preserving interdisciplinary scientific data[J]. Earth science informatics, 2010, 3(1/2): 101-110.
[23] LAGOZE C, PAYETTE S, SHIN E, et al. Fedora: an architecture for complex objects and their relationships[J]. International journal on digital libraries, 2006, 6(2): 124-138.
[24] HEDGES M, HASAN A, BLANKE T. Curation and preservation of research data in an iRODS data grid [EB/OL]. [2017-04-10]. http://ieeexplore.ieee.org/abstract/document/4426919/.
[25] LYON L. Dealing with data: roles, rights, responsibilities and relationships. consultancy report[EB/OL]. [2017-04-10]. http://opus.bath.ac.uk/412/.
[26] BRANDT D S. Librarians as partners in e-research Purdue University Libraries promote collaboration[J]. College & research libraries news, 2007, 68(6): 365-396.
[27] LAGE K, LOSOFF B, MANESS J. Receptivity to library involvement in scientific data curation: a case study at the University of Colorado Boulder[J]. portal: libraries and the academy, 2011, 11(4): 915-937.
[28] BRAZMA A, HINGAMP P, QUACKENBUSH J, et al. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data[J]. Nature genetics, 2001, 29(4): 365-371.
[29] BALL C A, SHERLOCK G, PARKINSON H, et al. Standards for microarray data[J]. Science, 2002, 298(5593): 539-539.
[30] SANSONE S-A, ROCCA-SERRA P, FIELD D, et al. Toward interoperable bioscience data[J]. Nature genetics, 2012, 44(2): 121-126.
[31] HOWE D, COSTANZO M, FEY P, et al. Big data: the future of biocuration [J]. Nature, 2008, 455(7209): 47-50.
[32] GOOD B M, CLARKE E L, LOGUERCIO S, et al. Building a biomedical semantic network in Wikipedia with Semantic Wiki Links[J]. Database, 2012, 2012: bar060.
[33] ASHBURNER M, BALL C A, BLAKE J A, et al. Gene Ontology: tool for the unification of biology [J]. Nature genetics, 2000, 25(1): 25-34.
[34] APWEILER R, BAIROCH A, WU C H, et al. UniProt: the universal protein knowledgebase [J]. Nucleic acids research, 2004, 32(S1): D115-D119.
[35] BAHL A, BRUNK B, CRABTREE J, et al. PlasmoDB: the Plasmodium genome resource. a database integrating experimental and computational data [J]. Nucleic acids research, 2003, 31(1): 212-215.
[36] GOERTSCHES R H, HECKER M, KOCZAN D, et al. Long-term genome-wide blood RNA expression profiles yield novel molecular response candidates for IFN-β-1b treatment in relapsing remitting MS [J]. Pharmacogenomics, 2010, 11(2): 147-161.
[37] 周喜斌 , 韩文静 , 陈晶 , 等 . 几种 QSAR 建模方法在化学中的应用与研究进展 [J]. 计算机与应用化学 , 2011, 28(6): 761-765.
[38] FOURCHES D, MURATOV E, TROPSHA A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research [J]. Journal of chemical information and modeling, 2010, 50(7): 1189-1204.
[39] TROPSHA A. Best practices for QSAR model development, validation, and exploitation [J]. Molecular informatics, 2010, 29(6/7): 476-488.
[40] FU X, WOJAK A, NEAGU D, et al. Data governance in predictive toxicology: a review[J]. Journal of cheminformatics, 2011, 3(1): 24.
[41] WILLIAMS A J, EKINS S. A quality alert and call for improved curation of public chemistry databases [J]. Drug discovery today, 2011, 16(17): 747-750.

知识管理论坛，2017(3):201−213
DOI：10.13266/j.issn.2095−5472.2017.022

[42] SCHENA M. DNA microarrays: a practical approach[M]. Oxford:Oxford University Press, 1999.
[43] TONG W, CAO X, HARRIS S, et al. ArrayTrack--supporting toxicogenomic research at the US Food and Drug Administration National Center for Toxicological Research [J]. Environmental health perspectives, 2003, 111(15): 1819.
[44] PRYOR G, DONNELLY M. Skilling up to do data: whose role, whose responsibility, whose career? [J]. International journal of digital curation, 2009, 4(2): 158-170.
[45] TIBBO H R. Placing the horse before the cart: conceptual and technical dimensions of digital curation [J]. Historical social research, 2012,37(3):187-200.
[46] LYON L. The informatics transform: re-engineering libraries for the data decade [J]. International journal of digital curation, 2012, 7(1): 126-138.
[47] HSWE P. Data management services in libraries [EB/OL]. [2017-04-10]. http://pubs.acs.org/doi/pdf/10.1021/bk-2012-1110.ch007.
[48] CHOUDHURY G S. Case study in data curation at Johns Hopkins University [J]. Library trends, 2008, 57(2): 211-220.
[49] DELSERONE L M. At the watershed: preparing for research data management and stewardship at the University of Minnesota Libraries [J]. Library trends, 2008, 57(2): 202-210.

Research on Mapping the Knowledge Domain of Digital Curation ——A Bibliometric Study of Web of Science (1990-2016)

Yu Chenlin[1,2]
[1]National Science Library, Chinese Academy of Sciences, Beijing 100190
[2]University of Chinese Academy of Sciences, Beijing 100049

Abstract: [Purpose/significance] Digital Curation (DC) is an important part of research data management in e-Science environment. By systematically combing the research progress and proposing some issues worthy of further studying, this paper aims to provide a basis and reference for national research data management research. **[Method/process]** Documents relevant to data curation were retrieved from Web of Science database. With CiteSpace Ⅲ software based on the document co-citation analysis method, this paper drew the knowledge map of international digital curation. In terms of research-based subject distribution, research institutions, researchers and knowledge of the four dimensions, the related contents were analyzed and summarized with the content analysis method. **[Result/conclusion]** International digital curation research began in 2000. Now it has entered a mature period, with a specific research disciplines, institutions and groups. The research's knowledge base is data description, integration and association, data maintenance in the scientific research process and value-added activities, data management stakeholders and service innovation models of library services.

Keywords: digital curation data management research data knowledge mapping